中央研究院資訊科學研究所 20 週年慶特刊

http://www.iis.sinica.edu.tw/page/aboutus/20thAnniversary/i05.html

謝清俊先生之簡介與訪談簡要

謝清俊先生於 2001 年 8 月自本所退休，但仍為兼任研究員，續主持本所文獻處理實驗室的工作，並協助本院規劃與執行國科會自 2002 年 1 月起為期五年的「數位典藏國家型科技計畫」，為計畫辦公室主任。謝先生於 1983 年至中央研究院，擔任資訊科學研究所研究員。於同年受前院長吳大猷先生囑託籌備計算中心，並於 1983 至 1990 年兼計算中心籌備處主任至主任一職。謝先生的研究工作涵蓋中文資訊處理、中文圖書館自動化、古籍資料庫、數位典藏、以及資訊技術對社會影響。於謝先生退休前夕（2001 年 7 月 26 日），我們與他進行了一次訪談。茲摘錄部份訪談內容如下，以饗讀者。

問：你在中文資訊處理的基礎工作，從早期的中文輸入與輸出系統，字與詞的頻率統計，到近期的中文缺字處理系統。請你就「字形」（glyph）處理方面，說明目前的進展以及未來的展望？

中文輸入與輸出涉及的是中文的構字問題。我是花了一些時間來瞭解文字學，把文字學中用古文所寫的東西，用科學、數學的方式表達出來，也就是等於說把文字學中的科學結構用現在的數學、邏輯語言表達出來，一旦轉換成現在的數學語言，計算機就可以處理這些文字。

這也涉及了中文字形交換碼的問題。中文字形交換碼的問題在於中文與外國語文的結構並不相同。外國語文的結構是詞（word），詞之下就是字母（alphabet）。字母是符號，本身沒有意義，經過一個次序組織出詞之後，才有意義。詞的意義多半是由音延伸而來。在這個情形下，外國語文的好處是很簡潔，字母是一個封閉集合（closed set），數量有限，所以數學結構表達起來非常容易。中文則有豐富的語意成分，是以知識來表達所構成的對象，而不是以表音為主。中文文字就是將古代的日常生活用圖像表達出來，裡頭也同時包含了語音與語意兩個部分。在這個基礎上，中文的文字並非是一個封閉集合，而是一個開放集合（open set）。

外國人並不懂中文的這套結構，因此當他們開始處理中文編碼時，直接就把中文字跟字母對應起來。將中文字跟字母對等是有問題的，因為中文字是開放集合，把開放集合納入封閉集合的結構中，有些字一定處理不了。有些字處理不了好像不是什麼大問題，但卻是非常嚴重的文化歧視。一直到今日，中文交換碼的最大問題都還是想把開放集合當作封閉集合來處理，所以不管那套系統都會有中文缺字的問題。

其實中文字是有封閉集合的，就是當初造字時的元件（components），古代曾經將這些元件分為聲母與形母兩類，加起來約有幾千個。這個統計與分類的工作在漢朝許慎做說文解字時就統計過，說文解字歸納出五百四十個部首，但其中仍然有問題。一方面是資料量太大；許慎處理了一萬一千多個字，分析這麼大量的資料難免有錯誤之處。另一方面則受到秦始皇焚書的影響，也就是說資料收集不易。

我開始做這個工作時，就是想承續這些文字學的努力，把中國構字的法則找出來，希望將中文構字用現代科學、數學的語言表達出來。最初我是在交大任職時處理這項工作，不過那時候受到電腦處理容量的限制，例如說記憶體只有八千位元（8k bytes）。分析字根時得盡量節省計算資源，因此必須忽略一些細節。到了院裡的時候，我就把這個工作再重新整理一遍，將過去忽略的細節重新補足，以忠實中國文字學的結構。最近我們發展出來的系統裡，大概包含了 1200 個字根，這些字根就是中國字形的最基本結構，字根與字根間則透過一些規則來結合成文字。另外，現在的系統也考慮到使

用者的方便運用，因此我們將字根擴展到約 4000 個元件，讓任何人使用時只須做一個層次的分析，只要考慮一個組合運算，就能直觀的知道一個字是如何組成。

我們用這個系統來處理缺字，這個系統最終將是中文交換碼的下一代，因為所有現在的交換碼都是封閉集合，而我們的系統是一套產生系統（production system）。透過這個系統我們已經可以處理六萬多個字，而現行的任何系統都無法處理六萬多字。這當中，文字是用構字式來表達，而構字式是唯一的。只要是構字式不同，就是不同的字（雖然一個字可能有不同的構字式，但都可以經過計算機的化約歸納到一個標準式），字和字就可以比對了。

透過這套系統，可以做中文字的交換。它和傳統編碼的方式完全是兩套不同的思路，但可以附加在目前任何編碼系統上，都沒有問題。現行的系統只要加上一個後處理（post process）運算來識別沒有碼的那些缺字，無須更改原來的編碼。譬如大五碼（big 5），可以加上構字式的系統，其中所有的字根都用大五碼的編碼，但缺字就用構字式來表達。目前何建明的實驗室做了一個網站，根據這些構字式就可以到網站上去下載那些字形。不只如此，這套系統也可以延伸到處理日文、韓國、越南的漢字，甚至於中國各個朝代的文字，在時、空上都沒有問題。

問：你曾經參與「中文資訊交換碼」（Chinese Character Code for Information Interchange, CCCII）的制訂與推廣，並規劃與建立中央研究院圖書館自動化系統。請你就這方面，作一些回顧？「中文資訊交換碼」的前景為何？

中文資訊交換碼最初是由我領導的小組所制訂的。1979 年 11 月時，美國為了處理東亞文字，想要訂定一套標準碼，就由美國國會圖書館委託史丹佛大學，由 John Haeger 所負責的研究圖書館組織（Research Library Group, RLG）召集一個會議。那時國科會的處長王紀五要我去參加這個會。在那個會議中，我發現由於當時全世界只有一套日本的 JIS 交換碼可處理漢字。如果我們沒有一套中文資訊交換碼的話，可能日本的交換碼就會變成全部漢字的標準，我覺得這是個非常嚴肅的事。

那時我原本就在收集資料，1979 年教育部也剛好發表了一份 4808 個常用字的字集，使我們可以開始做中文資訊交換碼的工作，不過在召開 RLG 會議的時候，我們還沒有開始進行。而美國其實是非常急於要訂立一套標準的。因為在軍事以及其他的用途上，美國必須用計算機處理中、日、韓文等東方語文資料，其中最大的問題即在中文字的處理。他們打算在第二年（1980）三月在華盛頓召開亞洲研究學會年會時，決定採取哪套編碼系統。

為了避免中文系統採用日本漢字的編碼方式，在 RLG 會議時，我就說我們正在做中文編碼的工作。一回國，我就跟當時電機工程學會會長李國鼎先生報告，我對他說，這是很嚴重的問題，牽涉到我們的文化的問題。舉例來看，清朝末年編了一本電報號碼本，到那時還是中文字唯一的編碼標準，這個電報號碼本編得很糟糕，所以用起來很不方便。外國人在設計 ISO 646（ISO 7-bit coded character set for information interchange，字元集資訊交換碼）時，中國人沒有參與，所以計算機根本沒有辦法處理中文。這個標準如果不制訂的話。我們會對不起以後我們的子孫。

李國鼎聽了很贊同，就找了一些基金會，募了三百萬台幣，要我組織一個小組。因此我找了一些文字學家，如金祥恆、趙友培、潘重規以及當初的中央圖書館館長王振鵠等，以及一些圖書館方面的教授等。1979 年 12 月 25 日正式成立了「國字整理小組」，花了三個月的時間，將 4808 個字的字集編碼作出來。當時張仲陶教授和我一起做這個事，他那年沒有回家過年，就在台灣技術學院的計算中心幫我跑這些資料。我上飛機時，帶了幾本上去，裝訂的膠都還沒有乾，拿在手上還是軟的。

就是在這個情形下，我在 1980 年 3 月，到美國報告。我在上面報告說明時，底下有美國圖書館界的編碼專家、ANSI（美國國家標準局）的編碼專家、語言方面的專家等，考了我一個早上。這四個小時比我考博士論文還辛苦。之後，中午吃飯時，他們恭喜我說，決定採用我們這套系統，不用日本的系統。日本派了七、八個代表去，但最後日本的系統沒被採用，非常失望的回去了。

問：中文資訊交換碼為什麼沒有成為國家標準？

這是很辛酸的事情，大部分的事情已經都收錄在 1989 年出版的《國字整理小組十年》一書中。其中一個問題是，在做中文資訊交換碼時，我把簡體字也都收錄進來，很多人就罵我「與匪認同」。李國鼎把我找去，我很簡單的說：「我們要不要處理匪情資料？」他說要。我說：「我們不是說反攻大陸以後要把簡體字跟正體字做一個對照，要做很多文宣，這些工作需不需要計算機幫忙？」他說要。「所以我把簡體字放進去。」聽了之後，李國鼎就幫我把這些事都檔掉。

但後來有一些人，認為中文編碼不過是把中文排列起來給個編號而已，為什麼要給你做？他們批評說，CCCII 用三個 bytes 來做編碼是不對的，三個 bytes 太浪費空間，我們用兩個 bytes 就有 65536 個組合，使用其中一部份編一萬多字就夠了。實際上是他們並沒有搞清楚什麼是資料碼（data code），什麼是控制碼（control code），結果編出來的碼裡頭沒有控制碼。最後的結果是完全不能使用，後來修改時，還是把 CCCII 一步一步學過去。當時的一些科技大老要求我就這兩套編碼做妥協，說新編一個系統，一部份採取我的結構，另一部份採取他的結構。我直接拒絕，因為兩套系統有很多不相容之處。在我拒絕之後，他們就弄出了一套 CNS（國家標準交換碼），但是 CNS 至今並沒有人使用。今日用的是 big5，而非 CNS。至於 big5，是民間發展出來的，業界要發展中文系統，但 CNS 不能用，又不能用謝清俊的，因為使用謝清俊的就會得罪那些大老。只好弄一個 big5 出來。

CCCII 一直維護到三、四年前張仲陶教授過世以前，之後我也不再做維護了。我認為，CCCII 從 1979 年發展出來，經過 20 多年後，階段性的任務已經完成，沒有必要再維持下去。而且現在這個缺字系統出來，就可以取代 CCCII 了。

問：中研院在古籍資料庫的工作，很早就開始了。請問你在古籍資料庫方面的工作經驗與心得？是哪些因素驅使你進行這些工作？

我離開台灣技術學院（編注：現台灣科技大學）到中研院的主要目的，就是想做古籍資料庫。當時外國可以看到很多文史方面的工作已經開始進行，中國卻沒有。1983 年

我到院裡來，就跟史語所毛漢光先生講好到這裡來做古籍的事。但是來不到半年就被吳大猷先生抓去做計算中心的籌備工作。當時中研院的計算環境很差，除了資訊所有一些電腦外，大概就只有植物所有一台迷你電腦（mini computer），大概是 HP 2100 系列的，其他沒有一個所有電腦。吳大猷先生跟我說得語重心長。他說他雖然不懂電腦，但知道以後研究不用電腦的話，研究院的學術地位在世界上會受到非常大的挑戰。這句話說的很沈重的，因為他一直說他不需要用電腦做研究，他只需要一枝鉛筆一張紙就夠了。吳大猷先生做研究確實是如此，但是對年輕人就不一樣，他也瞭解電腦對研究院的發展很重要。

這件事我到今天仍然不瞭解是誰推薦我去的，因為之前我根本不認識吳大猷先生。我跟吳大猷先生說，我的想法是要盡全力幫忙文史的人。我說，自然科學和生命科學我們不用照顧他們，機器給他們用就好了，但是文史方面必須要有人照顧他們。接了計算機中心主任之後，我就利用這個機會，把我想做的二十五史的資料做出來。從那時一直到我卸任之前，計算機中心的資源幾乎百分之七十在支持文史，一直到現在我也覺得這個策略沒有錯，因為電腦買了，對自然科學、生命科學的人來說，他們很容易的就可以使用，但是文史的非有人帶不可。

在計算機中心時，我就開始做古籍電子化的工作，研究院開始做電子文獻時，外國剛起步不到兩年。我們從 1985 開始做二十五史全文資料庫，國外是 1984 年年底開始有一些全文資料庫出現，所以事實上我們做全文資料庫是滿早的，確實相當有前瞻性。比起我們的計算機科學，技術跟國外至少五到十年的差距來說，是少得多了。

從文史的角度來思考，其實和從計算機科學的角度來思考是不一樣的。舉個例來說，訂二十五史全文資料庫的規格時，我手下有一些人，拿了一些國外的論文來跟我說，國外的全文資料庫的檔案結構都是一頁文稿一個檔案，你為什麼要堅持一個段落做一個單位？為什麼一定要堅持要保留二十五史原書的段落、行數跟字數？我跟他們講道理講不通，他們從計算機的技術來看認為我是在找麻煩，但是我認為那是錯的。這是因為我認為結構分成好多種，版面結構是一種，文章內容結構是另外一種，我們必須要知道哪些資訊是做全文資料庫時必須保留下來的。這到今天都是一個好問題。

這個制訂出來規格其實是非常領先的，當時 SGML（Standard Generalized Markup Language，通用標示語言語法）正在發展，SGML 是 1986 正式發展出來的，我們在 1985 已經訂出了一個自己的標示語言，一直到今天計算機中心還在使用。後來上網路時，才將這套系統對應到 HTML（HyperText Markup Language，超文件標示語言）格式。我們那套標示語言和 SGML 不一樣。雖然 SGML 功能較強，但也比較複雜。我們的系統用拿來做二十五史也就夠了。